Back to the Table of Contents

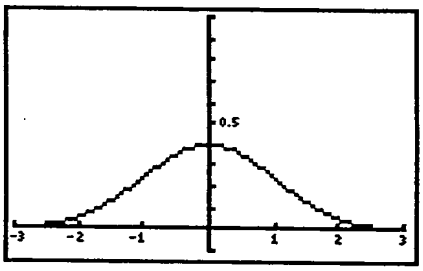
An Introduction to Statistics - Lesson 6

The Bell-shaped, Normal, Gaussian Distribution

Lesson Overview

The Bell-shaped, Normal, Gaussian Distribution

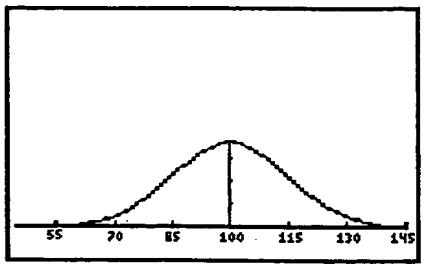
It can be shown under very general assumptions that the distribution of independent random errors of observation takes on a **normal** distribution as the number of observations becomes large. Although others were involved, Gauss was one of the first to characterize this distribution and hence it is often named after him. It is also shaped like a bell, hence yet another name. The term used in the title above is rather redundant, but serves to emphasize that the three are identical. You can graph this curve on your calculator as seen below by entering the following function: $y=e^{-x^2/2}/\sqrt{(2\pi)}$, where e is the transcendental number 2.71828... and π is the more familiar, but also transcendental number 3.14159.... The π in the formula only serves to **normalize** the total area under the curve. When we normalize something, we make it equal to some **norm** or standard, usually one (1).



The standard normal distribution

The height of the curve represents the probability of the measurement at that given distance away from the mean. The total area under the curve being one represents the fact that we are 100% certain (probability = 1.00) the measurement is somewhere. Technically, this is the **standard normal** curve which has $\mu=0.0$ and $\sigma=1.0$. Other applications of the normal curve do not have this restriction. For example, intelligence has often been cast, allbeit controversially, as **normally distributed** with $\mu=100.0$ and $\sigma=15.0$. This is represented below. Our function has been modified to $y=e^{-(x-\mu)^2/2\sigma^2}/(\sigma\sqrt{(2\pi)})$

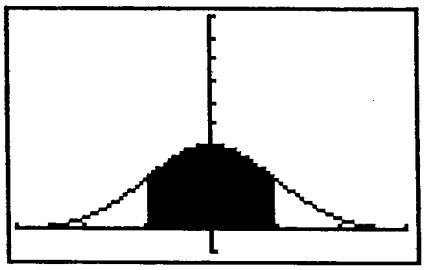
Normally distributed IQs



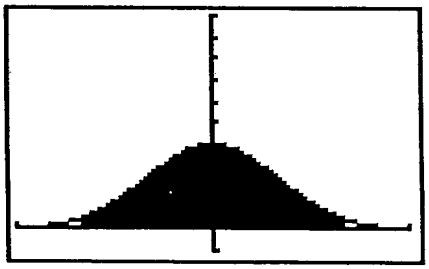
Other things which may take on a normal distribution include body temperature, shoe sizes, diameters of trees, *etc.* It is also important to note the **symmetry** of the normal curve. Some curves may be slightly distorted or truncated beyond certain limits, but still primarily conform to a "heap" or "mound" shape (see below). This is often an important consideration when analyzing data or samples taken from some unknown population.

The Empirical Rule

For a normally distributed data set, the **empirical rule** states that 68% of the data elements are within one standard deviation of the mean, 95% are within two standard deviations, and 99.7% are within three standard deviations. Graphically, this corresponds to the area under the curve as shown below for 1 and 2 standard deviations. The empirical rule is often stated simply as **68-95-99.7**. Note how this ties in with the range rule of thumb, by stating that 95% of the data usually falls within two standard deviations of the mean.



Data within 1σ (left) and 2σ (right)



The author usually claims an IQ of at least 145. We can see from the above information that this would put him about three standard deviations above the population mean (100+3•15=145). Hence, if we accept the hypothesis that IQs are normally distributed, 99.85% of the population would have a lower IQ and 0.15% a higher one. Please especially note that if 99.7% of the population is within three standard deviations of the mean, the remaining 0.3% is distributed with half beyond three standard deviations below the mean and the other half beyond three standard deviations above the mean. This is a result of the symmetry (due to the fact that x is squared, it matters not if it is positive or negative) of the curve. In practical terms, in a population of 250,000,000; 249,625,000 would have an IQ lower than 145 and 375,000 would have an IQ higher. Because of the small area of these regions, they are often referred to as **tails**. Depending on the circumstances, we may be interested in **one tail** or **two tails**.

Several societies exist which cater to individuals with high IQs. Some specific examples would be MENSA, Triple Nine, Mega, *etc.*.

Another important charateristic of this distribution is that it is of **infinite extent**. In practical terms, IQs below 0 (6.67σ) or above 210 (7.33σ) (ceiling scores such as Marilyn Vos Savant's are difficult to interpret) do not occur. A recently popularized manufacturing goal has been termed Six Sigma. One

would think this would corresponds with about 3.4 defects per billion, but their web site implies it is 200 per million. A typically good company operates at less than four sigma or 99.997% perfect. This corresponds closer to 32 defects per million. If you have ever purchased a "lemon" (a colloquialism for bad car, perhaps one built on a Monday) you can appreciate such striving for perfection. Other similar examples would be the large increase in errors related to perscription drugs being dispensed or the case of the Florida patient who had the wrong leg amputated.

Chebyshev's Theorem

Chebyshev (1821-1894) was a preeminent Russian mathematician who primarily worked on the theory of prime numbers, although his writings covered a wide range of subjects. One of those subjects was probability and his theorem applies to any data set, not only normally distributed data sets. His theorem states that the portion of any set of data within K standard deviations of the mean is always at least $1-1/K^2$, where K may be any number greater than 1.

For K=2, we see that $1-1/2^2=1-1/4=3/4$, which is 75% of the data must always be within two standard deviations of the mean.

For K=3, we see that $1-1/3^2=1-1/9=8/9$, which is about 89% of the data must always be within three standard deviations of the mean.

If we consider the data set 50, 50, 50, and 100, we will discover that the sample standard deviation (s) is 25, and the upper score falls exactly at $2s$ above the rest. However, since the mean is 62.5, it is well within $2s$. Added 5 more scores of 50 we find the mean is now 55.6 and the standard deviation now 16.7. We see that two standard deviations above the mean now extends to 88.9 and we have one data point outside that, but within three standard deviations. The general concept of being able to find the mean of a data set and determine how much of it is within a certain distance (number of standard deviations) of the mean is an important one which we will continue in the next lesson.

Note: here is an example of a data set with k=2 and only 75% of the data within the proscribed limits. It comes to us from Hogg and Craig (1978, p. 70) in "Introduction to Mathematical Statistics". (5th ed.) via the AP STAT list server on May 31, 2000. Let the discrete random variable x have probabilities 1/8, 6/8, 1/8 at the points x= -1, 0, 1 respectively. μ=0 and σ^2=1/4. If k=2, then $1/k^2$=1/4 and we thus attain the bound given by Chebyshev's inequality.

BACK HOMEWORK ACTIVITY CONTINUE

- e-mail: calkins@andrews.edu
- voice/mail: 269 471-6629/ BCM&S Smith Hall 106; Andrews University; Berrien Springs,
- **classroom:** 269 471-6646; Smith Hall 100/**FAX:** 269 471-3713; MI, 49104-0140
- home: 269 473-2572; 610 N. Main St.; Berrien Springs, MI 49103-1013
- URL: http://www.andrews.edu/~calkins/math/webtexts/stat06.htm